Numer Algor (2018) 77:1117–1139
DOI 10.1007/s11075-017-0355-4

ORIGINAL PAPER

A numerical method for stationary shock problems with monotonic solutions

Relja Vulanović[1] · **Thái Anh Nhan**[2]

Received: 19 February 2017 / Accepted: 29 May 2017 / Published online: 3 June 2017

Abstract Numerical methods are considered for singularly perturbed quasilinear problems having interior-shock solutions. It is shown that the direct discretization on a layer-adapted mesh is ineffective for these problems. A special method is proposed for the case when the solution is monotonic: the problem is transformed by interchanging the dependent and independent variables, and it is then discretized on a uniform mesh. The method is analyzed both theoretically and numerically. It is shown that it can be effective, but that it is not entirely without problems. An approach for improving the method is suggested.

Keywords Quasilinear boundary value problem · Singular perturbation · Interior shock · Finite differences

Mathematics Subject Classification (2010) 65L10 · 65L11 · 65L12 · 65L20

✉ Relja Vulanović
rvulanov@kent.edu

Thái Anh Nhan
anhan@ohlone.edu

[1] Department of Mathematical Sciences, Kent State University at Stark, 6000 Frank Ave. NW, North Canton, OH 44720, USA

[2] Department of Mathematics, Ohlone College, 43600 Mission Blvd., Fremont, CA 94539, USA

1 Introduction

We consider the problem of finding a $C^2[0,1]$-solution to the singularly perturbed boundary-value problem

$$-\varepsilon\frac{d^2u}{dx^2}-b(u)\frac{du}{dx}+c(u)=0,\quad x\in(0,1),\quad u(0)=A,\quad u(1)=B, \tag{1}$$

where ε is a small positive parameter, b and c are sufficiently smooth functions, and A and B are given constants. It is assumed throughout the paper that the following condition is satisfied:

$$\frac{d}{du}c(u)\geq c_*>0,\quad u\in\mathbb{R}. \tag{2}$$

The condition guarantees that the problem has a unique solution, [11].

We are exclusively interested here in the case when the solution, which we denote by u_ε, is strictly monotonically increasing. Throughout the paper we assume that

$$u_\varepsilon'(x)>0,\quad x\in[0,1], \tag{3}$$

and, necessarily, that $A<B$. According to [12, Lemma 2.1], (3) holds true if in addition to (2) we have

$$c(0)=0,\quad A<0<B. \tag{4}$$

The problem (1)–(2) is a challenging problem to solve numerically when u_ε has one or more interior layers. This happens under certain conditions (see [7, 12] for details) and the interior layers are located around the points p_ε such that $b(u_\varepsilon(p_\varepsilon))=0$. The exact value of p_ε is not known in general, but when $\varepsilon\to 0$, p_ε approaches the point x_*, where the solution of the corresponding reduced problem (problem (1) with $\varepsilon=0$) is discontinuous (has a shock). Points x_* can be determined, [7, 12].

Not even the numerical methods specialized for singular perturbation problems can resolve the interior layer for problems of type (1). This is because the corresponding interior layer of the discrete problem is shifted from the position where the layer of u_ε is located. An analogous situation can be observed in the shifted position of the soliton when the Korteweg-de Vries equation is solved numerically, [6]. Therefore, ε-uniform pointwise accuracy is very hard to achieve. The special layer-adapted meshes, like Shishkin's, are of no help here. This is why we want to explore here another, very unique, approach which utilizes the fact that u_ε is monotonically increasing. We call this approach the "inversion method" because we interchange the variables x and u in (1). This results in the "inverted problem,"

$$\varepsilon\left(\frac{1}{x'}\right)'-c(u)x'+b(u)=0,\quad u\in(A,B),\quad x(A)=0,\quad x(B)=1, \tag{5}$$

where $'=d/du$. Let x_ε be the solution of (5). It may happen that x_ε has no layers and then it suffices to discretize (5) on a uniform mesh.

Because of the difficulties mentioned above, ε-uniform numerical methods are often constructed for problems whose solution only simulates the interior-shock behavior. For instance, a linear problem of this kind is considered in [17] and a nonlinear one is analyzed in [5]. In these modifications of (1), the position of the interior layer is known in advance and is fixed in the sense of not depending on the numerical

solution. Another special case of (1) treated numerically is the boundary-layer case, in which $b(0) = 0$ and $u(0) = 0$, [16, 24, 27, 28].

The general case, but with $b(u) = u$, is considered in [15, 25]. The analysis is simpler when $b(u) = u$ because there is no more than one interior layer. The corresponding two-dimensional problem is dealt with in [21]. The three papers have in common that they analyze the solution of the problem (1) by considering its behavior separately on the left and right sides of the layer. The error of the approximate solution obtained in [25] contains an $\ln(1/\varepsilon)$-factor, so strictly speaking, this method is not uniform in ε. The analysis in [15] produces an error estimate which includes a $\mathcal{O}(|p_\varepsilon - q|/\varepsilon)$-term, where q is an approximate location of the interior layer ($q = x_*$ is used in the numerical experiments there). This shows how sensitive the direct discretization is with respect to our ability to pinpoint where the interior layer is located. An intricate algorithm for capturing the location of the layer is proposed by Shishkin in [21]. It involves $\mathcal{O}(N^{3/2})$ operations, where N is the number of mesh steps in each spatial direction, to obtain ε-uniform accuracy of the numerical solution of the order $\mathcal{O}(N^{-1/5} \ln^{1/2} N)$. In [22], Shishkin analyzes an analogous parabolic problem using the same approach. Some numerical experiments with Shishkin's method for the parabolic problem are provided in [18].

As opposed to the approaches described above, the inversion method only requires assumptions (2) and (3), and it does not need a special procedure for locating the layer(s). In Section 3, we introduce a discretization scheme for the inverted problem (5) and we prove under these general assumptions that the discrete problem has a unique monotonically increasing solution. Convergence uniform in ε is not proved; it is analyzed through numerical experiments in Section 4. We only consider test problems with solutions which have exactly one layer. The simplest problem of this kind is the Lagerstrom-Cole model problem ([8, p. 56], [9, p. 86], [14, p. 167]),

$$-\varepsilon \frac{d^2u}{dx^2} - u\frac{du}{dx} + u = 0, \quad x \in (0,1), \quad u(0) = A, \quad u(1) = B, \tag{6}$$

with appropriate conditions on A and B. Whereas the numerical results for (6) are mostly satisfactory, a slight shift in the position of the layer is still present when the method is applied to more complicated problems. In Section 5, we propose a method of improving the results, which is based on the value of x_*, but in a way that is different from those used in [15, 25].

We start off by using (6) in Section 2 to illustrate numerically that the direct discretization cannot resolve the interior layer effectively, and we conclude the paper by offering some final remarks in Section 6.

2 The inadequacy of the direct discretization

Problems of type (1) are usually solved numerically by discretizing the corresponding conservation form,

$$-\varepsilon \frac{d^2u}{dx^2} - \frac{d}{dx}a(u) + c(u) = 0, \quad x \in (0,1), \quad u(0) = A, \quad u(1) = B, \tag{7}$$

where

$$a(u) = \int_0^u b(t)dt.$$

The Engquist-Osher scheme [19] is one of the most often used schemes for solving (7) numerically. For its construction, we need the functions

$$a^{\pm}(u) = \int_0^u b^{\pm}(t)dt, \quad b^+ = \frac{1}{2}(b + |b|), \quad \text{and} \quad b^- = \frac{1}{2}(b - |b|). \tag{8}$$

Consider the discretization mesh with points $0 = x_0 < x_1 < x_2 < \cdots < x_{N-1} < x_N = 1$ and let u_i be the ith component of the numerical solution; $u_i \approx u_\varepsilon(x_i)$. Let also $h_i = x_i - x_{i-1}$, $i = 1, 2, \ldots, N$, and $\hbar_i = (h_i + h_{i+1})/2$, $i = 1, 2, \ldots, N-1$. The Engquist-Osher discretization of (7) is

$$-\varepsilon D''u_i - D^-a^-(u_i) - D^+a^+(u_i) + c(u_i) = 0, \quad i = 1, 2, \ldots, N-1, \tag{9}$$

where $u_0 := A$, $u_N := B$, and

$$D''u_i := \frac{1}{\hbar_i}\left(\frac{u_{i+1} - u_i}{h_{i+1}} - \frac{u_i - u_{i-1}}{h_i}\right),$$

$$D^-u_i := \frac{u_i - u_{i-1}}{\hbar_i}, \quad D^+u_i := \frac{u_{i+1} - u_i}{\hbar_i}.$$

The scheme is an upwind scheme for quasilinear problems and it is stable uniformly in ε, [12, 19]. On the uniform mesh, the upwind scheme cannot produce ε-uniform numerical results even for linear problems, let alone quasilinear ones. However, whereas special, layer-adapted meshes (like those of Bakhvalov or Shishkin types, [10, 20]) enable ε-uniform convergence for linear and boundary-layer quasilinear problems, such meshes do not work well for quasilinear problems with interior layers. We demonstrate this below.

We tested the direct discretization on the Lagerstrom-Cole problem (6) under the conditions that guarantee the presence of a unique shock at $x_* = (1 - A - B)/2$. These conditions are $B - A > 1$ and $B, -A \in [0, 1]$, and the reduced solution is

$$u_r = \begin{cases} x + A & \text{if } 0 \le x < x_*, \\ x - 1 + B & \text{if } x_* < x \le 1, \end{cases}$$

see [7–9, 12, 14]. It is shown in [12] that the Engquist-Osher scheme, when applied to the reduced problem, yields a numerical solution which approximates u_r with $\mathcal{O}(h)$-accuracy at all mesh points except for the two points surrounding the point x_*.

Because (4) is satisfied, the solution u_ε is monotonically increasing and there is a unique point p_ε such that $b(u_\varepsilon(p_\varepsilon)) = 0$; in this case $u_\varepsilon(p_\varepsilon) = 0$. The interior layer is located around p_ε, but since this point is not known, we used x_* as its approximation. We did the same in the numerical experiments of Section 4.

We solved the discrete problem (9) both on a uniform mesh and on a Shishkin mesh dense around x_*, which is defined as follows. Let

$$\tau = \min\{\eta, 2\varepsilon \ln N\}, \quad \eta = \frac{1}{2}\min\{x_*, 1 - x_*\}.$$

The mesh consists of three parts: a fine uniform mesh with $N/2$ mesh steps in the interval $[x_* - \tau, x_* + \tau]$ and two coarse uniform meshes, each with $N/4$ mesh steps,

one in the interval $[0, x_* - \tau]$ and another in $[x_* + \tau, 1]$. All numerical experiments presented here are for $\varepsilon = 10^{-6}$ and $N = 64$.

The nonlinear system of the discrete problem was solved by Newton's method with the initial guess formed by the values on the straight line between the points $(0, A)$ and $(1, B)$. The iterations were stopped when the maximum norm of the difference between two successive iterations dropped below the user-prescribed tolerance of 10^{-9}.

Figure 1 shows the graph of the numerical solution in the case when $A = -\frac{1}{2}$ and $B = 1$, which gives $x_* = \frac{1}{4}$. It is easily observed that the layer is not resolved well. The numerical solution has its own layer, which is in this case shifted to the left of x_*. This is what Herman and Knickerbocker [6] call a *numerically induced phase shift* in the position of the soliton, occurring when the Korteweg-de Vries equation is solved by the Zabusky-Kruskal scheme. In our case, the numerically induced shift places the layer where the mesh is not dense, and the fine mesh, around $x_* = \frac{1}{4}$, gives a cluster of 33 points above the right branch of the reduced solution u_r, see Fig. 2. This is the inadequacy of the Shishkin mesh (or any "layer-adapted" mesh) for interior-shock problems. The results are in fact even worse than what we can get using the uniform mesh (see Fig. 3), which, of course, cannot resolve the layer either.

Satisfactory results can only be obtained when both continuous and numerical solutions are centrally symmetric with respect to the point $(\frac{1}{2}, 0)$, like when $-A = B = 1$, giving $x_* = \frac{1}{2}$. Figure 4 shows this situation and illustrates what is meant by a "well-resolved layer." This is in a striking contrast to Fig. 1.

3 The discretization of the inverted problem

Recall that $A < B$. Let Ω^N be the discretization mesh with points $u_i^N = u_i = A + ih$, $i = 0, 1, \ldots, N$, where $h = (B - A)/N$. By x^N, y^N, etc., we denote mesh functions on $\Omega^N \setminus \{A, B\}$. Any mesh function x^N is identified with the corresponding

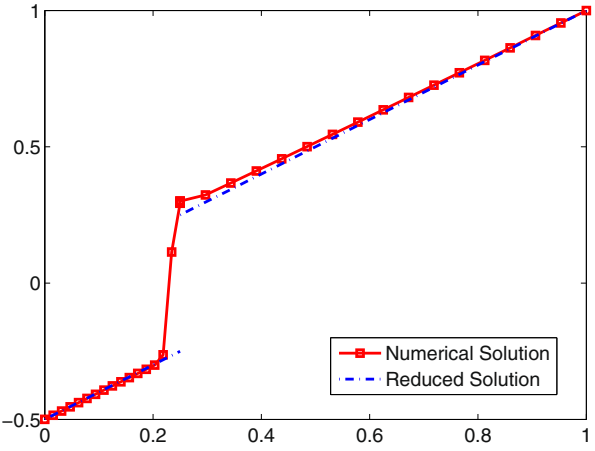

Fig. 1 Numerical solution of Eq. 6 with $A = -\frac{1}{2}$, $B = 1$, discretized on the Shishkin mesh

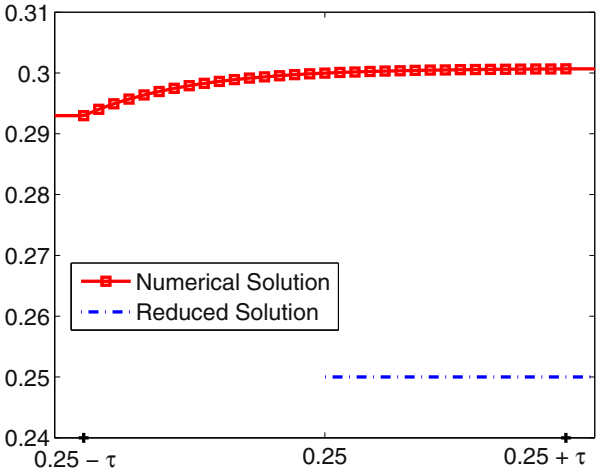

Fig. 2 A zoomed-in portion of the numerical solution presented in Fig. 1

$\mathbb{R}^{N-1}$ column-vector, $x^N = [x_1, x_2, \ldots, x_{N-1}]^T$. For simplicity, the superscript N is removed from mesh points and mesh-function components unless the value of N needs to be emphasized. We formally set $x_0 := 0$ and $x_N := 1$. Let $e^N := [1, 1, \ldots, 1]^T$. We are particularly interested in the monotonically increasing mesh functions; they belong to the set

$$X^N := \left\{ x^N \mid 0 < x_1 < x_2 < \cdots < x_{N-1} < 1 \right\}.$$

We shall also use

$$\bar{X}^N := \left\{ x^N \mid 0 \le x_1 \le x_2 \le \cdots \le x_{N-1} \le 1 \right\}.$$

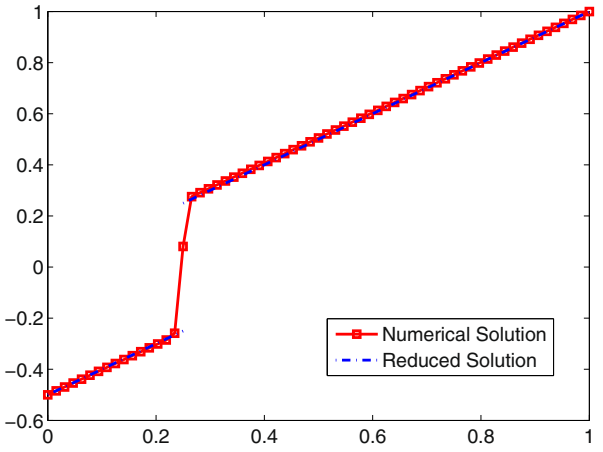

Fig. 3 Numerical solution of Eq. 6 with $A = -\frac{1}{2}$, $B = 1$, discretized on a uniform mesh

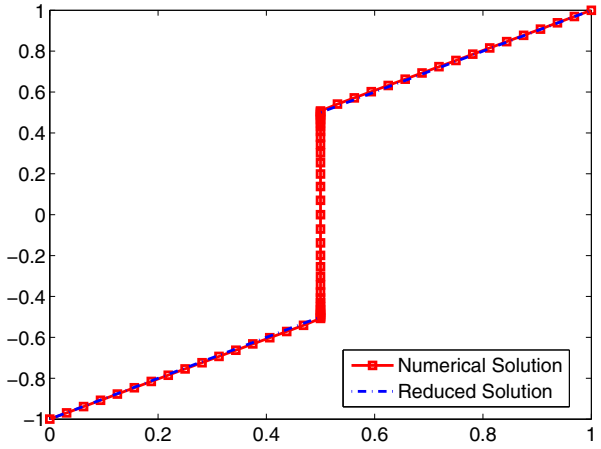

Fig. 4 Numerical solution of Eq. 6 with $A = -1$, $B = 1$, discretized on the Shishkin mesh

Let $\Delta^+ x_i = x_{i+1} - x_i$ and $\Delta^- x_i = \Delta^+ x_{i-1} = x_i - x_{i-1}$, and define the norms

$$\left\|x^N\right\|_\infty = \max_{1\le i\le N-1}\left|x_i\right| \quad \text{and} \quad \left\|x^N\right\|_1^h = h\left\|x^N\right\|_1, \quad \text{where} \quad \left\|x^N\right\|_1 = \sum_{i=1}^{N-1}\left|x_i\right|.$$

The matrix norm induced by the vector norm $\|\cdot\|_1$ (or $\|\cdot\|_1^h$) is also denoted by $\|\cdot\|_1$.

For any $C[A, B]$-function g, we use g^N to indicate the discretization of g on $\Omega^N\backslash\{A, B\}$. In particular, x_ε^N is the discretization on $\Omega^N\backslash\{A, B\}$ of the exact solution x_ε to the problem (5). We simply write g_i for $g(u_i)$. Let $g_{i\pm1/2} = g(u_i \pm h/2)$. We also write $g^\pm$ in the sense of $b^\pm$ in (8).

We discretize the inverted problem (5) on Ω^N as follows:

$$T^N x_i := \varepsilon\left(\frac{1}{\Delta^+ x_i} - \frac{1}{\Delta^- x_i}\right) - D'[c]x_i = -\hat{b}_i, \qquad i = 1, 2, \ldots, N-1, \tag{10}$$

where

$$D'[c]x_i := \frac{1-s_i}{2}\cdot\frac{c_{i-1/2}}{h}\Delta^- x_i + \frac{1+s_i}{2}\cdot\frac{c_{i+1/2}}{h}\Delta^+ x_i$$

and

$$\hat{b}_i = \frac{1-s_i}{2}b_{i-1/2} + \frac{1+s_i}{2}b_{i+1/2}$$

with

$$s_i = \operatorname{sign} c_i = \begin{cases} 1 & \text{if } c_i > 0, \\ 0 & \text{if } c_i = 0, \\ -1 & \text{if } c_i < 0. \end{cases}$$

The above scheme for the $c(u)x'$-term is the second-order midpoint upwind scheme. If $c_i > 0$, the scheme reduces to

$$D'[c]x_i = c_{i+1/2}\frac{x_{i+1} - x_i}{h},$$

which approximates $(cx')_{i+1/2}$ with second-order accuracy. Note that then $c_{i+1/2} > 0$ because of (2). On the other hand, if $c_i < 0$, then $c_{i-1/2} < 0$ and

$$D'[c]x_i = c_{i-1/2}\frac{x_i - x_{i-1}}{h}$$

is a second-order approximation of $(cx')_{i-1/2}$. If $c_i = 0$ (which because of (2) cannot happen more than once), $D'[c]$ is a transition scheme which averages the above midpoint upwind schemes. The way $D'[c]$ changes is accompanied with the corresponding changes in $\hat{b}_i$. This means that the reduced problem outside the layer is solved with second-order accuracy.

The main result of the paper follows.

Theorem 1 *Assume the condition* (2)*. Then, the following is true:*

(a) *The discrete problem* (10) *has a solution* $\tilde{x}^N \in X^N$.

(b) *The discrete operator* T^N *is* ε*-uniformly stable in* X^N*. More precisely, the stability inequality*

$$\left\|x^N - y^N\right\|_1^h \le \frac{2}{c_*}\left\|T^N x^N - T^N y^N\right\|_1^h$$

is satisfied for any two mesh functions $x^N, y^N \in X^N$*. Therefore, the solution* $\tilde{x}^N$ *is unique in* X^N.

(c) *For some positive constant* K*,*

$$\left\|\tilde{x}^N - x_\varepsilon^N\right\|_1^h \le Kh,$$

where K *may depend on* ε*, but is independent of* h.

Proof (a) The first part of the theorem is proved in several steps.

1° We first consider an auxiliary linear system, defined using a fixed vector $y^N \in \bar{X}^N$ and a positive constant σ:

$$L^N[y^N]x_i = \sigma y_i, \quad i = 1, 2, \ldots, N-1, \tag{11}$$

where the linear operator $L^N[y^N]$ is given by

$$L^N[y^N]x_i := \varepsilon\left(\Delta^- x_i - \Delta^+ x_i\right) - \Delta^- y_i \Delta^+ y_i D'[c]x_i + D[b, y^N]x_i + \sigma x_i$$

and

$$D[b, y^N]x_i := \hat{b}_i^+ \Delta^+ y_i \Delta^- x_i + \hat{b}_i^- \Delta^- y_i \Delta^+ x_i.$$

Let $M = [m_{ij}]$ be the matrix of the system (11). It is a tridiagonal matrix and its entries satisfy

$$m_{i,i-1} = -\varepsilon + \Delta^- y_i \Delta^+ y_i \frac{c_{i-1}^-}{h} - \hat{b}_i^+ \Delta^+ y_i < 0, \quad i = 2, 3, \ldots, N-1,$$

$$m_{i,i+1} = -\varepsilon - \Delta^- y_i \Delta^+ y_i \frac{c_{i+1}^+}{h} + \hat{b}_i^- \Delta^- y_i < 0, \quad i = 1, 2, \ldots, N-2,$$

and finally, upon formally setting $m_{10} := 0$ and $m_{N-1,N} := 0$,

$$m_{ii} \ge \sigma - m_{i,i-1} - m_{i,i+1}, \quad i = 1, 2, \ldots, N-1,$$

where the equality holds true for $i = 2, 3, \ldots, N-2$. Therefore, M is an L-matrix satisfying $Me^N \geq \sigma e^N$. This implies that M is an inverse-monotone matrix. Equivalently, $L^N[y^N]$ is an inverse-monotone operator for which the following stability inequality holds true for any two vectors v^N and w^N:

$$\left\|v^N - w^N\right\|_\infty \leq \frac{1}{\sigma}\left\|L^N[y^N]\left(v^N - w^N\right)\right\|_\infty. \tag{12}$$

Because of this, the system (11) has a unique solution, which we denote by $x^N(y^N)$.

2° We now show that $x^N(y^N)$ is in $\bar{X}^N$. First, $0 \leq x_i(y^N) \leq 1$, $i = 1, 2, \ldots, N-1$, because $L^N[y^N]$ is inverse monotone and we have that

$$L^N[y^N]0 = 0 \leq \sigma y_i = L^N[y^N]x_i(y^N) \leq \sigma = L^N[y^N]1, \quad i = 1, 2, \ldots, N-1,$$

(keep in mind that $x_0(y^N) = y_0 = 0$ and $x_N(y^N) = y_N = 1$). Then, we consider the differences $d_i := \Delta^+ x_i(y^N)$ and we want to show that $d_i \geq 0$, $i = 0, 1, \ldots, N-1$. We already have that

$$d_0 = x_1(y^N) \geq 0 \ \text{ and } \ d_{N-1} = 1 - x_{N-1}(y^N) \geq 0. \tag{13}$$

The differences d_i satisfy the system

$$\Lambda^N d_i = \sigma(y_{i+1} - y_i), \quad i = 1, 2, \ldots, N-2, \tag{14}$$

where

$$\begin{aligned}\Lambda^N d_i &:= L^N[y^N]x_{i+1}(y^N) - L^N[y^N]x_i(y^N)\\ &= m_{i,i-1}d_{i-1} + (\sigma - m_{i+1,i} - m_{i,i+1})d_i + m_{i+1,i+2}d_{i+1}.\end{aligned}$$

Let $P = [p_{ij}] \in \mathbb{R}^{N-2,N-2}$ be the matrix of the system (14). It is obvious that P is an L-matrix. We also have

$$\left(P^T e^N\right)_i = p_{i-1,i} + p_{ii} + p_{i+1,i} = \sigma, \quad i = 2, 3, \ldots, N-3,$$

and moreover,

$$\left(P^T e^N\right)_1 = p_{11} + p_{21} = \sigma - m_{12} > \sigma$$

and

$$\left(P^T e^N\right)_{N-2} = p_{N-2,N-2} + p_{N-3,N-2} = \sigma - m_{N-1,N-2} > \sigma.$$

This means that $P^T e^N \geq \sigma e^N$ and, therefore, P^T is an inverse-monotone matrix, and so is P. Thus, the operator Λ^N is also inverse monotone. Then, since (14) yields that $\Lambda^N d_i \geq 0 = \Lambda^N 0$ and since (13) holds true, it follows that $d_i \geq 0$, i.e., $x^N(y^N) \in \bar{X}^N$.

3° We can now define the mapping G, $G : \bar{X}^N \to \bar{X}^N$,

$$Gy^N = x^N, \quad \text{where } \ x^N = x^N(y^N), \quad \text{that is, } \ L[y^N]x^N = \sigma y^N.$$

To show that G is continuous, consider $Gy^{j,N} = x^{j,N}$, $j = 1, 2$. We use (12) to get

$$\begin{aligned}\left\|Gy^{1,N} - Gy^{2,N}\right\|_\infty &= \left\|x^{1,N} - x^{2,N}\right\|_\infty \\ &\le \frac{1}{\sigma}\left\|L^N[y^{1,N}]x^{1,N} - L^N[y^{1,N}]x^{2,N}\right\|_\infty \\ &\le \frac{1}{\sigma}\left\|\sigma y^{1,N} - \sigma y^{2,N}\right\|_\infty \\ &\quad + \frac{1}{\sigma}\left\|L^N[y^{2,N}]x^{2,N} - L^N[y^{1,N}]x^{2,N}\right\|_\infty \\ &\le \kappa\left\|y^{1,N} - y^{2,N}\right\|_\infty\end{aligned}$$

with some positive constant κ, $\kappa > 1$. This constant may depend on ε and h, but that is irrelevant here. We have that G is continuous and we can use the Brouwer fixed-point theorem to conclude that G has a fixed point $\tilde{x}^N \in \bar{X}^N$. The fixed point satisfies $L^N[\tilde{x}^N]\tilde{x}^N = \sigma\tilde{x}^N$ and this system reduces to

$$\varepsilon\left(\Delta^-\tilde{x}_i - \Delta^+\tilde{x}_i\right) - \Delta^-\tilde{x}_i\Delta^+\tilde{x}_i\left(D'[c]\tilde{x}_i - \hat{b}_i\right) = 0, \quad i = 1, 2, \ldots, N-1. \tag{15}$$

4° We now show that $\tilde{x}^N \in X^N$. Suppose $\tilde{x}_{j-1} = \tilde{x}_j$ for some j, i.e., $\Delta^-\tilde{x}_j = 0$. Then the equations in (15) imply that $\tilde{x}_j = \tilde{x}_{j+1}$ and $\tilde{x}_{j-2} = \tilde{x}_{j-1}$ and the conditions $\tilde{x}_0 = 0$ and $\tilde{x}_N = 1$ cannot be satisfied. Since $\tilde{x}^N \in X^N$, the system (15) can be rewritten as

$$T^N\tilde{x}_i = -\hat{b}_i, \quad i = 1, 2, \ldots N-1.$$

This completes the proof of part (a) of the theorem.

(b) For any $x^N \in X^N$, let $F = [f_{ij}]$ be the Fréchet derivative at x^N of the operator T^N. We have that F is a tridiagonal matrix with the entries

$$\begin{aligned}f_{i,i-1} &= -\frac{\varepsilon}{\left(\Delta^-x_i\right)^2} + \frac{1-s_i}{2}\cdot\frac{c_{i-1/2}}{h} < 0, \quad i = 2, 3, \ldots, N-1, \\ f_{ii} &= \frac{\varepsilon}{\left(\Delta^+x_i\right)^2} + \frac{\varepsilon}{\left(\Delta^-x_i\right)^2} - \frac{1-s_i}{2}\cdot\frac{c_{i-1/2}}{h} + \frac{1+s_i}{2}\cdot\frac{c_{i+1/2}}{h} > 0, \\ i &= 1, 2, \ldots, N-1, \\ f_{i,i+1} &= -\frac{\varepsilon}{\left(\Delta^+x_i\right)^2} - \frac{1+s_i}{2}\cdot\frac{c_{i+1/2}}{h} < 0, \quad i = 1, 2, \ldots, N-2.\end{aligned}$$

Our next goal is to show that

$$F^T e^N \ge \frac{c_*}{2}e^N, \tag{16}$$

which implies the desired stability inequality. For $i = 2, 3, \ldots, N-2$, we have

$$\left(F^T e^N\right)_i = f_{i-1,i} + f_{ii} + f_{i+1,i} = \frac{2 + s_i - s_{i+1}}{2}\cdot\frac{c_{i+1/2}}{h} + \frac{s_i - s_{i-1} - 2}{2}\cdot\frac{c_{i-1/2}}{h}.$$

If $s_{i-1} \geq 0$, then $s_i = s_{i+1} = 1$ and $c_{i-1/2} > 0$, and it follows that

$$\left(F^T e^N\right)_i \geq \frac{c_{i+1/2}}{h} - \frac{c_{i-1/2}}{h} \geq c_*.$$

The same inequality holds true for $s_{i+1} \leq 0$. If $s_{i-1} = -1$ and $s_{i+1} = 1$, then s_i can have any of the three possible values. For instance, if $s_i = -1$, we have

$$\left(F^T e^N\right)_i = -\frac{c_{i-1/2}}{h} > \frac{c_i}{h} - \frac{c_{i-1/2}}{h} \geq \frac{c_*}{2}.$$

The remaining cases, $s_i = 0$ and $s_i = 1$, can be treated in the same way. It can also be shown that

$$\left(F^T e^N\right)_1 = f_{11} + f_{21} > \frac{c_*}{2},$$

as well as

$$\left(F^T e^N\right)_{N-1} = f_{N-2,N-1} + f_{N-1,N-1} > \frac{c_*}{2}.$$

Therefore, (16) is satisfied.

(c) This result follows from the stability inequality in part (b) and the fact that T^N is an $\mathcal{O}_\varepsilon(h)$ scheme for the inverted continuous problem (5). □

Remark 1 We comment here on the different techniques used in the proof of Theorem 1.

The use of the Brouwer fixed theorem to prove the existence of a solution to a discretization of a quasilinear singularly perturbed boundary-value problem is due to Zadorin [28]. The problem considered there is of a non-turning-point type and the scheme is an exponentially fitted one. This technique is used in [23] for direct discretizations of quasilinear problems with monotonic solutions. It is adapted here to the inverted problem, particularly in the way the discrete operator L^N is constructed. The proof of part (a) 2° uses the same technique as in [26] and part (a) 4° is like in [12, 19].

Remark 2 The scheme (10) is not the only one for which Theorem 1 can be proved. In general, the inverted problem (5) has to be discretized in its original form, without switching to the conservation form. This is because the corresponding linear operator L^N in step (a) 1° of the proof of Theorem 1 has to be stable in the maximum norm. On the other hand, the scheme should also be stable in norm $\|\cdot\|_1^h$ because of part (b) of Theorem 1.

The regular upwind scheme for discretizing (5) is like (10), but with

$$\bar{D}'[c]x_i := \frac{c_i^-}{h}\Delta^- x_i + \frac{c_i^+}{h}\Delta^+ x_i$$

instead of $D'[c]x_i$ and b_i instead of $\hat{b}_i$. However, Theorem 1 cannot be proved for this scheme. The difficulty is in part (b) of the proof and it occurs at the columns $j-1$ and j of the Fréchet derivative of the discrete operator, where j is such an index that

$c_{j-1} < 0 \le c_j$. This can be rectified by adding a $\mathcal{O}(h^2)$-term to the scheme at the points u_{j-1} and u_j:

$$\bar{T}^N x_i := \varepsilon \left(\frac{1}{\Delta^+ x_i} - \frac{1}{\Delta^- x_i} \right) - \bar{D}'[c]x_i - \gamma_i \big(\Delta^+ x_i - \Delta^- x_i \big) = -b_i, \quad i = 1, 2, \ldots N-1,$$

where

$$\gamma_i = \begin{cases} \gamma & \text{if } i = j-1, j, \\ 0 & \text{otherwise,} \end{cases}$$

with γ a fixed constant in $(0, c_*)$. (If $c_j = 0$, γ_{j-1} may be also set equal to 0.)

Another possible modification of the regular upwind scheme, which satisfies Theorem 1, is

$$\tilde{T}^N x_i := \varepsilon \left(\frac{1}{\Delta^+ x_i} - \frac{1}{\Delta^- x_i} \right) - \tilde{D}'[c]x_i = -b_i, \quad i = 1, 2, \ldots N-1,$$

where

$$\tilde{D}'[c]x_i := \frac{1 - s_i}{2} \cdot \frac{c_{i-1}}{h} \Delta^- x_i + \frac{1 + s_i}{2} \cdot \frac{c_{i+1}}{h} \Delta^+ x_i.$$

The $c(u)x'$-term can also be discretized using the following scheme:

$$\check{T}^N x_i := \varepsilon \left(\frac{1}{\Delta^+ x_i} - \frac{1}{\Delta^- x_i} \right) - \check{D}'[c]x_i = -b_i, \quad i = 1, 2, \ldots N-1,$$

where

$$\check{D}'[c]x_i = \frac{1}{2h} \big[c_i \big(x_{i+1} - x_{i-1} \big) + \Gamma \big(\Delta^+ x_i - \Delta^- x_i \big) \big].$$

and $|c(u)| \le \Gamma$, $u \in [A, B]$. This is similar to the Lax-Friedrichs scheme, [19].

Remark 3 Under the conditions of Theorem 1 we have

$$\sum_{i=1}^{N-1} \big| u_\varepsilon(\tilde{x}_i) - u_i \big| \le \tilde{K},$$

where $\tilde{K}$ is a positive constant, which may depend on ε, but is independent of h. This follows from part (c) of Theorem 1 and the fact that $u_i = u_\varepsilon(x_\varepsilon(u_i))$. Therefore, the values u_i approximate those of $u_\varepsilon(\tilde{x}_i)$ with first-order accuracy in the discrete L^1 norm. It cannot be concluded from here that the inversion method produces ε-uniform pointwise accuracy. We leave it to numerical experiments to see whether this can be achieved.

4 Numerical results

We experimented with the scheme (10) and the schemes mentioned in Remark 2. The scheme (10) produced the most accurate results. We only present these results here.

All examples satisfy the conditions (2) and (4), so it is guaranteed that u_ε is strictly monotonically increasing.

Since Newton's method works fine for the direct discretization (9), we also considered it for solving the nonlinear system (10) representing the discretization of the inverted problem. We wanted to use a general initial guess that can work for all test problems. The values on the straight line between $(A, 0)$ and $(B, 1)$ represented a natural choice. However, with this initial iteration, Newton's method (which is well-known for its sensitivity to the initial guess) only converged to a monotonic solution when ε was close to 1. To enable the convergence of Newton's method for all values of ε that we considered in our experiments, we applied ε-extrapolation, that is, we combined Newton's method with ε-iterations. We experimented with sequences of ε-values decreasing either arithmetically or geometrically. Although our intention was not to find a procedure that would require the smallest number of iterations, the geometric sequence generally performed better in that sense than the arithmetic one and it is the only sequence we present below.

Let $\varepsilon^* \in (0, 1]$ be an ε-value for which we have obtained the solution of the discrete problem (10) and let $\varepsilon_* \in (0, \varepsilon^*)$ be the value of ε for which we want to produce new numerical results. We define a sequence of ε-values,

$$\varepsilon_i = \varepsilon_{i-1} \sqrt[k]{\frac{\varepsilon_*}{\varepsilon^*}}, \quad i = 1, 2, \ldots, k, \quad \varepsilon_0 = \varepsilon^*,$$

so that $\varepsilon_k = \varepsilon_*$. All tables of results for the inversion method in this section are created by solving first the discrete problem for $\varepsilon = \varepsilon_0 = 1$ using Newton's method with the initial guess created by the straight line between the points corresponding to the boundary conditions. This numerical solution serves as the initial guess for Newton's method applied to the discrete problem with $\varepsilon = \varepsilon_1$. The procedure continues in the same manner, i.e., the numerical solution for $\varepsilon = \varepsilon_{i-1}$ is the initial guess for Newton's method used to solve the discrete problem with $\varepsilon = \varepsilon_i$. All Newton's iterations $x^{N,m}$ are calculated until $\|x^{N,m} - x^{N,m-1}\| \le \mathrm{TOL}$, where TOL is a prescribed tolerance. The result for $\varepsilon = \varepsilon_k = \varepsilon_*$ is recorded and used as the initial guess for the next smaller value of ε in the table. Typical choices in our numerical experiments were $\mathrm{TOL} \le 10^{-14}$ and $k = 5$.

While it is the inverted problem (5) that is solved numerically, we are more interested in how accurately the solution u_ε of the original problem (1) is calculated, rather than the solution x_ε of the inverted problem. This is why we want to estimate the errors

$$E_\varepsilon^N := \max_{1 \le i \le N-1} \left| u_\varepsilon(x_i) - u_i \right|. \tag{17}$$

We also calculate the numerical order of convergence of the scheme using

$$\mathrm{Ord}_\varepsilon^N := \log_2 E_\varepsilon^N - \log_2 E_\varepsilon^{2N}$$

and the numerical order of ε-uniform convergence as

$$\mathrm{Ord}^N := \min_\varepsilon \mathrm{Ord}_\varepsilon^N.$$

The first test problem is the linear boundary-layer problem

$$-\varepsilon u'' - u' + u = 0, \quad x \in (0, 1), \quad u(0) = -1, \quad u(1) = 1. \tag{18}$$

The solution has a layer in the neighborhood of $x = 0$. Of course, the inversion method is not created for such simple problems, but the exact solution u_ε is known here and we can find the exact values of the error (17). The results are presented in Table 1. The errors do not increase as ε decreases, which indicates that the method is ε-uniformly convergent. The values of Ord^N are close to 1, as expected from a first-order scheme. The table also contains the number of iterations, $\mathrm{Iter}_\varepsilon$, defined as

$$\mathrm{Iter}_\varepsilon = \max_N \mathrm{Iter}_\varepsilon^N,$$

where $\mathrm{Iter}_\varepsilon^N$ is the total number of Newton iterations for all ε-iterations between two consecutive ε-values shown in the table. The values of $\mathrm{Iter}_\varepsilon^N$ stabilize as N increases and change very little for $\varepsilon \le 10^{-3}$ (it is known in general that the number of Newton iterations is independent of N, [1]). The value of $\mathrm{Iter}_\varepsilon$ for the greatest ε in any table includes all iterations from the initial value of $\varepsilon = 1$.

After the above boundary-layer problem, we consider two test problems with solutions that have an interior layer and no other layers. In these problems, $b(0) = 0$, so that the value $u = 0$ corresponds to the x-value where the layer is located.

The first such problem is the Lagerstrom-Cole problem (6) with different values of A and B that guarantee the presence of an interior layer. In general, when $B \le 1$, $A \ge -1$, and $B - A > 1$, the shock is at $x_* = (1 - A - B)/2$ and the asymptotic solution can be given as (see [8])

$$\tilde{u}_\varepsilon(x) := \begin{cases} x + A & \text{if } 0 \le x < x_* + \dfrac{1}{\theta}\varepsilon \ln \varepsilon, \\ \theta \tanh \dfrac{\theta(x - x_*)}{2\varepsilon} & \text{if } |x - x_*| \le -\dfrac{1}{\theta}\varepsilon \ln \varepsilon, \\ x - 1 + B & \text{if } x_* - \dfrac{1}{\theta}\varepsilon \ln \varepsilon < x \le 1, \end{cases}$$

where

$$\theta = \frac{B - A - 1}{2} > 0.$$

For this problem, instead of using (17), we estimate the error by

$$\tilde{E}_\varepsilon^N := \max_{1 \le i \le N-1} \left| \tilde{u}_\varepsilon(x_i) - u_i \right|$$

Table 1 Errors E_ε^N for the linear boundary-layer problem (18)

$-\log_{10}\varepsilon$	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$	$\mathrm{Iter}_\varepsilon$
1	2.04e–02	1.05e–02	5.34e–03	2.69e–03	1.35e–03	49
2	2.54e–02	1.33e–02	6.84e–03	3.47e–03	1.75e–03	47
3	2.70e–02	1.40e–02	7.22e–03	3.69e–03	1.87e–03	44
4	2.72e–02	1.41e–02	7.27e–03	3.73e–03	1.89e–03	40
5	2.73e–02	1.41e–02	7.27e–03	3.73e–03	1.89e–03	40
6	2.73e–02	1.41e-02	7.27e–03	3.74e–03	1.89e–03	40
Ord^N	0.95	0.96	0.96	0.98		

Table 2 Errors $\tilde{E}_\varepsilon^N$ for the Lagerstrom-Cole problem (6) with $A = -1$, $B = 1$ ($x_* = \frac{1}{2}$)

$-\log_{10}\varepsilon$	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$	Iter_ε
5	2.28e–02	1.25e–02	6.57e–03	3.41e–03	1.71e–03	297
6	2.30e–02	1.26e–02	6.66e–03	3.51e–03	1.80e–03	52
7	2.31e–02	1.26e–02	6.67e–03	3.52e–03	1.81e–03	48
8	2.31e–02	1.26e–02	6.67e–03	3.52e–03	1.81e–03	44
9	2.31e–02	1.26e–02	6.67e–03	3.52e–03	1.81e–03	40
$\widetilde{\text{Ord}}^N$	0.87	0.92	0.92	0.96		

and the order of convergence by the corresponding $\widetilde{\text{Ord}}_\varepsilon^N$ and $\widetilde{\text{Ord}}^N$, calculated analogously to Ord_ε^N and Ord^N.

We present the results for two cases of the problem (6). The first case is with $A = -1$ and $B = 1$, when the shock is at $x_* = \frac{1}{2}$ and the solution u_ε is symmetric about $(\frac{1}{2}, 0)$. The other case is $A = -\frac{1}{2}$ and $B = 1$, giving $x_* = \frac{1}{4}$ and an asymmetric solution. Our numerical experiments revealed that 0 has to be a mesh point; otherwise, a numerically induced shift is still present and the results are not satisfactory. Therefore, when $A = -1$ and $B = 1$, N needs to be even, and when $A = -\frac{1}{2}$ and $B = 1$, N needs to be divisible by 3. As Tables 2 and 4 show, the results behave similarly to those in Table 1.

The errors in Table 2 can be compared to the errors presented in Table 3, which result from the direct Engquist-Osher discretization on the Shishkin mesh (see Section 2; one such numerical solution is shown in Fig. 4). For the values of N considered in Table 3, the numerical orders of convergence still do not show the influence of $\ln N$ factors, which are typically present in the errors when the Shishkin mesh is used. We can see from Tables 2 and 3 that the inverse discretization outperforms the direct one in terms of accuracy.

Of course, when it comes to the Lagerstrom-Cole problem (6) with asymmetric solution, the inversion method is incomparably better. This can be confirmed by taking a look of Tables 4 and 5 and the graphs in Figs. 5 and 1. Whereas the errors for the inversion method in the asymmetric case (Table 4) behave similarly to the errors in the symmetric case (Table 2), those for the direct method in the asymmetric case do not even indicate convergence when N increases (Table 5).

If $B/(-A)$ is a rational number, we can find two positive integers N_1 and N_0 such that $B/(-A) = N_1/N_0$. Then, $N = N_0 + N_1$ guarantees that 0 is a mesh point. We

Table 3 Errors $\tilde{E}_\varepsilon^N$ for the Lagerstrom-Cole problem (6) with $A = -1$, $B = 1$ ($x_* = \frac{1}{2}$) solved by the Engquist-Osher scheme on the Shishkin mesh

$-\log_{10}\varepsilon$	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$
5	4.40e–02	2.28e–02	1.19e–02	6.21e–03	3.26e–03
6	4.39e–02	2.28e–02	1.18e–02	6.11e–03	3.14e–03
7, 8, 9	4.39e–02	2.28e–02	1.18e–02	6.10e–03	3.13e–03
$\widetilde{\text{Ord}}^N$	0.95	0.94	0.94	0.93	

Table 4 Errors $\tilde{E}_\varepsilon^N$ for the Lagerstrom-Cole problem (6) with $A = -\frac{1}{2}$, $B = 1$ ($x_* = \frac{1}{4}$)

$-\log_{10}\varepsilon$	$N = 30$	$N = 60$	$N = 120$	$N = 240$	$N = 480$	Iter_ε
5	1.58e–02	9.29e–03	4.91e–03	2.53e–03	1.22e–03	329
6	1.62e–02	9.53e–03	5.11e–03	2.70e–03	1.40e–03	61
7	1.63e–02	9.57e–03	5.13e–03	2.72e–03	1.42e–03	55
8	1.63e–02	9.58e–03	5.14e–03	2.73e–03	1.42e–03	50
9	1.63e–02	9.58e–03	5.14e–03	2.73e–03	1.42e–03	45
$\widetilde{\text{Ord}}^N$	0.77	0.90	0.91	0.94		

experimented with other asymmetric cases with $B/(-A)$ rational and got results similar to the case $A = -\frac{1}{2}$, $B = 1$. However, if $B/(-A)$ is irrational, there is no uniform mesh containing 0 as a mesh point and a numerically induced shift occurs. A slightly nonuniform mesh is needed to make 0 a mesh point when $B/(-A)$ is irrational and it is possible to construct a nonuniform generalization of the scheme (10) and to prove for it a result analogous to Theorem 1. We experimented with the Lagerstrom-Cole problem (6) with $A = -\frac{\sqrt{2}}{2}$ and $B = 1$, giving $x_* = \frac{\sqrt{2}}{4}$. We tried two types of nonuniform meshes, those that are slightly nonuniform around 0 and those that uniform around 0 and become nonuniform away from the layer. Neither approach gave satisfactory results. We can still report that a good rational approximation of an irrational $B/(-A)$ can produce reasonably accurate results on a uniform mesh, although the uniformity in ε cannot be entirely preserved as ε decreases. It is also interesting to point out that in this case greater accuracy cannot be achieved by doubling the previous values of N_0 and N_1, but by increasing N_0 and N_1 so that N_1/N_0 becomes a more accurate approximation of $B/(-A)$.

The asymptotic solution does not represent u_ε well for greater values of ε and this is why Tables 2–5 only contain results for $\varepsilon \le 10^{-5}$. When ε is greater, errors can be estimated using the double-mesh principle, described, for example, in [4]. We tested the principle on the direct Engquist-Osher discretization (9) of the Lagerstrom-Cole problem (6) on the Shishkin mesh, see Section 2. The results correctly indicate first-order accuracy for the symmetric problem with $A = -1$, $B = 1$, and the inadequacy of the method for the asymmetric problem with $A = -\frac{1}{2}$ and $B = 1$. For the inverse discretization, the double-mesh principle is applied as follows. Once x^N and x^{2N} are calculated for a particular ε, a piecewise linear interpolant $u^{I,2N}$ is created using the points $\left(x_i^{2N}, u_i^{2N}\right)$, $i = 0, 1, \ldots, 2N$. Then, the values of $u^{I,2N}(x_i^N)$ are compared to u_i^N and the following error is found:

$$E_\varepsilon^{I,N} = \max_{1 \le i \le N-1} \left| u_i^N - u^{I,2N}(x_i^N) \right|.$$

Table 5 Errors $\tilde{E}_\varepsilon^N$ for the Lagerstrom-Cole problem (6) with $A = -\frac{1}{2}$, $B = 1$ ($x_* = \frac{1}{4}$) solved by the Engquist-Osher scheme on the Shishkin mesh

$-\log_{10}\varepsilon$	$N = 30$	$N = 60$	$N = 120$	$N = 240$	$N = 480$
5	5.03e–01	4.88e–01	4.82e–01	4.81e–01	4.82e-01
6,7,8,9	5.04e–01	4.88e–01	4.82e–01	4.82e–01	4.84e–01

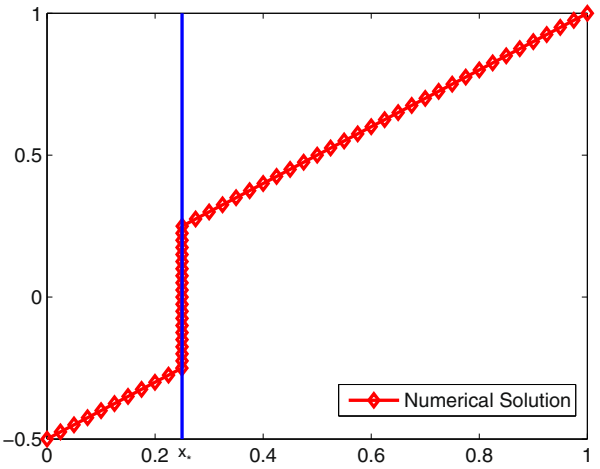

Fig. 5 Numerical solution of Eq. 6 with $A = -\frac{1}{2}$, $B = 1$ ($x_* = \frac{1}{4}$), $\varepsilon = 10^{-6}$, obtained by the inversion method with $N = 60$

Results for the asymmetric Lagerstrom-Cole problem are given in Table 6. The table also shows the corresponding numerical order of convergence $\mathrm{Ord}^{I,N}$, which is analogous to Ord^{N}. As N increases, $\mathrm{Ord}^{I,N}$ increases as well, but the values are still well below 1 for these values of N. The errors stabilize when $\varepsilon \to 0$, indicating the uniformity in ε. When compared to Table 4, the errors are smaller, but the orders of convergence are lower.

The second interior-layer problem is from [12]:

$$-\varepsilon u'' - \frac{u}{1+u}u' + u = 0, \quad x \in (0,1), \quad u(0) = A, \quad u(1) = B. \tag{19}$$

Table 6 Errors $E_{\varepsilon}^{I,N}$ for the Lagerstrom-Cole problem (6) with $A = -\frac{1}{2}$, $B = 1$ ($x_* = \frac{1}{4}$)

$-\log_{10}\varepsilon$	$N = 30$	$N = 60$	$N = 120$	$N = 240$
1	1.13e–03	6.18e–04	3.22e–04	1.64e–04
2	3.49e–03	1.98e–03	1.07e–03	5.52e–04
3	4.48e–03	3.12e–03	1.86e–03	1.02e–03
4	5.17e–03	3.36e–03	2.11e–03	1.18e–03
5	5.42e–03	3.53e–03	2.17e–03	1.21e–03
6	5.50e–03	3.58e–03	2.19e–03	1.22e–03
7	5.54e–03	3.60e–03	2.19e–03	1.22e–03
8	5.55e–03	3.60e–03	2.20e–03	1.22e–03
9	5.55e–03	3.60e–03	2.20e–03	1.23e–03
$\mathrm{Ord}^{I,N}$	0.52	0.67	0.84	

The reduced solution is

$$u_r = \begin{cases} u_L := (A+1)e^x - 1 & \text{if } 0 \le x < x_*, \\ u_R := (B+1)e^{x-1} - 1 & \text{if } x_* < x \le 1, \end{cases}$$

where the location x_* of the shock is found from the equation $a(u_L(x)) = a(u_R(x))$,

$$x_* = \ln\left(\ln\frac{B+1}{A+1} - 1\right) - \ln\left(\frac{B+1}{e} - A - 1\right).$$

Like in [12], we take $A = -\frac{1}{2}$ and $B = 2$, which gives $x_* = 0.271282$. In this problem, we again have $b(u) = 0$ when $u = 0$ and we make 0 a mesh point by using N divisible by 5.

The graph of the numerical solution for $\varepsilon = 10^{-6}$ and $N = 60$, presented in Fig. 6, looks quite acceptable, but the naked eye cannot detect that the position of the numerical layer is in fact slightly shifted from x_*. This is shown in Fig. 7. At the same time, the errors calculated using the double-mesh principle are not satisfactory for smaller values of ε, see Table 7. These results can be compared to Table 8 which shows the errors of the direct discretization by the Engquist-Osher scheme on the Shishkin mesh with the fine part around $x_* = 0.271282$. Except for $\varepsilon = 0.1$, the errors for larger values of ε are smaller when the inversion method is used, but, generally speaking, both methods produce useless results. The graph obtained by the direct method is similar to what we have seen for the asymmetric Lagerstrom-Cole problem (Fig. 1). It is of small consolation that the inversion-method graph looks better.

To explain the difficulties we encountered when solving (19), we compare the problem to the Lagerstrom-Cole problem (6). The main difference between the two problems is that the former has nonlinear reduced solutions u_L and u_R, as opposed to the linear u_L and u_R of (6). This means that the scheme we use is practically exact

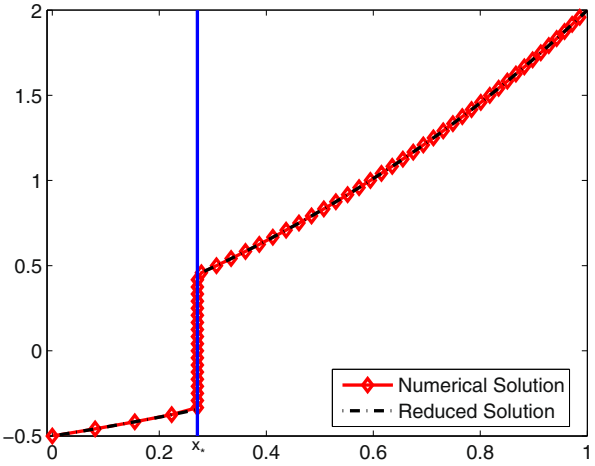

Fig. 6 Numerical solution of Eq. 19 with $A = -\frac{1}{2}$, $B = 2$, $\varepsilon = 10^{-6}$, obtained by the inversion method with $N = 60$

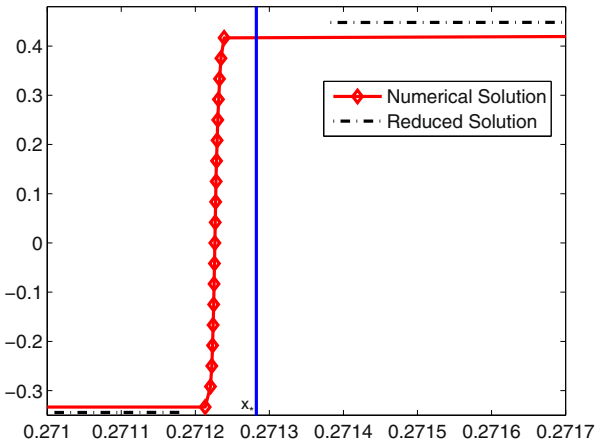

Fig. 7 A zoomed-in portion of the numerical solution presented in Fig. 6

outside the layer when applied to (6), but not when applied to (19). We attribute the difficulties with (19) to this fact.

5 Improving the results

We cannot be completely satisfied with the results of the inversion method for the Lagerstrom-Cole problem (6) when $B/(-A)$ is irrational and the results for the problem (19) are even worse, particularly for smaller values of ε. We now describe how it is possible to improve these results. Let $j = j(N)$ be such that $u_{j-1} < 0 \le u_j$. We are interested in the situation when $u_j > 0$. In our experiments with the Lagerstrom-Cole problem and $B/(-A)$ irrational, we noticed that a slight change in N (by one more point, for instance) typically causes the corresponding x_j values to be on different sides of x_*. Motivated by this, we consider two numerical solutions, x^{N_1} and x^{N_2}, where N_1 and N_2 are different but close. Let

$$\alpha_k = x^{N_k}_{j(N_k)} - x_*, \quad k = 1, 2,$$

Table 7 Errors $E_\varepsilon^{I,N}$ for the problem (19) with $A = -\frac{1}{2}$, $B = 2$

$-\log_{10}\varepsilon$	$N = 30$	$N = 60$	$N = 120$	$N = 240$
1	1.00e–02	5.26e–03	2.66e–03	1.33e–03
2	7.42e–03	4.39e–03	2.40e–03	1.35e–03
3	4.98e–02	5.98e–03	3.52e–03	1.90e–03
4	5.34e–01	4.89e–02	3.87e–03	5.97e–03
5	7.51e–01	4.79e–01	7.80e–02	7.07e–02
6	7.51e–01	7.50e–01	6.14e–01	5.62e–01
7, 8, 9	7.51e–01	7.50e–01	7.71e–01	7.81e–01

Table 8 Errors $E_\varepsilon^{I,N}$ for the problem (19) with $A = -\frac{1}{2}$, $B = 2$ solved by the the Engquist-Osher scheme on the Shishkin mesh

$-\log_{10}\varepsilon$	$N = 30$	$N = 60$	$N = 120$	$N = 240$
1	5.20e–03	2.81e–03	1.46e–03	7.44e–04
2	2.25e–01	1.11e–01	5.58e–02	2.74e–02
3	5.58e–01	4.21e–01	3.07e–01	2.35e–01
4	6.10e–01	5.80e–01	5.64e–01	5.32e–01
5–9	6.14e–01	5.90e–01	5.80e–01	5.74e–01

with $\alpha_1\alpha_2 < 0$. We see that for

$$\alpha = \frac{\alpha_2}{\alpha_2 - \alpha_1}$$

we get

$$\alpha x^{N_1}_{j(N_1)} + (1-\alpha)x^{N_2}_{j(N_2)} = x_*.$$

Based on this, we consider the linear combination

$$x^{N_1,N_2} := \alpha x^{I,N_1} + (1-\alpha)x^{I,N_2},$$

where x^{I,N_k} is the piecewise linear interpolant corresponding to the solution x^{N_k}. We take

$$x^{N_1,N_2}(u_i^{N_1}) = \alpha x_i^{N_1} + (1-\alpha)x^{I,N_2}(u_i^{N_1}) \tag{20}$$

as the new numerical solution instead of $x_i^{N_1}$, $i = 1, 2, \ldots, N_1 - 1$. The inequality $\alpha_1\alpha_2 < 0$ is desirable in this construction because it is equivalent to $0 < \alpha < 1$, which itself implies that x^{N_1,N_2} remains monotonically increasing.

The value of x_*, which is needed in the above approach, can be found from the equation $a(u_L(x)) = a(u_R(x))$, [12]. This equation does not involve ε and standard nonlinear solvers may be used to determine x_* with arbitrary accuracy.

When applying the linear combination (20) to the Lagerstrom-Cole problem, we used $N_1 = N$ and $N_2 = N + 1$, and got $\alpha_1\alpha_2 < 0$. The results are presented in Table 9. They are comparable to those in Tables 2 and 4.

As for the problem (19), we were able to find suitable values of N_1 and N_2 (those producing $\alpha_1\alpha_2 < 0$) only for smaller values of ε, but this is exactly where the need for improvement is greatest. We used $N_1 = N$ and $N_2 = N + 1$ and obtained the results presented in Table 10. The errors behave like in the corresponding part of

Table 9 Errors $\tilde{E}_\varepsilon^N$ of the linear combination (20) for the Lagerstrom-Cole problem (6) with $A = -\frac{\sqrt{2}}{2}$, $B = 1$ ($x_* = \frac{\sqrt{2}}{4}$)

$-\log_{10}\varepsilon$	$N = 30$	$N = 60$	$N = 120$	$N = 240$	$N = 480$
5	2.93e–02	1.33e–02	7.67e–03	3.37e–03	1.63e–03
6	2.94e–02	1.34e–02	7.83e–03	3.47e–03	1.74e–03
7	2.94e–02	1.35e–02	7.85e–03	3.48e–03	1.75e–03
8	2.94e–02	1.35e–02	7.85e–03	3.48e–03	1.76e–03
9	2.94e–02	1.35e–02	7.85e–03	3.48e–03	1.76e–03
$\widetilde{\text{Ord}}^N$	1.12	0.78	1.17	0.98	

Table 10 Errors $E_{\varepsilon}^{I,N}$ of the linear combination (20) for the problem (19) with $A = -\frac{1}{2}$, $B = 2$

$-\log_{10}\varepsilon$	$N = 30$	$N = 60$	$N = 120$	$N = 240$
6	2.38e–02	6.19e–03	4.22e–03	2.20e–03
7	2.39e–02	6.24e–03	4.11e–03	2.25e–03
8	2.39e–02	6.24e–03	4.10e–03	2.27e–03
9	2.39e–02	6.24e–03	4.10e–03	2.27e–03
$\mathrm{Ord}^{I,N}$	1.94	0.55	0.85	

Table 6. Regarding the greater values of ε, it should be mentioned that it is questionable whether the numerical layer should be placed at x_*. The location of the layer is at the point p_ε (recall that $b(u(p_\varepsilon)) = 0$) and x_* only approximates p_ε. This approximation is more accurate for smaller values of ε.

6 Conclusion

In this paper, we have introduced the inversion method, a very special numerical method for solving one-dimensional quasilinear interior-shock problems (1)–(2) in the case when they have strictly monotonic solutions. The monotonicity requirement comes from the main idea of the method, which is to interchange the independent and dependent variables and then to discretize the problem. For problems (1)–(2), there exists a convenient sufficient condition, (4), which guarantees that the solution is strictly monotonically increasing. The class of problems includes the Lagerstrom-Cole model problem, about which much has been written, [8, 9, 14]. Another related problem, which arises in applications, is the steady-state Burgers equation. Although the Burgers equation is of a different type since $c \equiv 0$, the corresponding boundary-value problem also has a monotonically increasing solution when $A < B$, [14, pp. 12–15]. Of course, it cannot be expected of application problems in general to satisfy (4), or to have strictly monotonic solutions (cf. the quasilinear application problems in [3] for instance). Although the scope of the inversion method is limited to one-dimensional problems with strictly monotonic solutions, we are motivated to consider it as an alternative to the direct discretization methods, which happen to be inadequate when applied to quasilinear interior-shock problems. As for the condition (4), it should be pointed out that it is not required in our analysis of the inversion method, since only (2) and $A < B$ are needed in the proof of Theorem 1. Therefore, the main idea of the inversion method applies to any one-dimensional problem if it is known that the problem has a strictly monotonic solution. We are not aware of some other simple condition like (4), but this information may come from the physical meaning of the problem or from preliminary numerical experiments. It should be kept in mind that any theoretical analysis of the method has to be adjusted to the specific problem, in the same way as it has been tailored here to problems (1)–(2).

Our numerical results have shown that the inversion method is generally better than the direct discretization. However, the inversion method is not without problems of its own. The numerically induced shift in the position of the layer, from

which the direct method suffers acutely, is still present in all problems solved by the inversion method, except the simplest Lagerstrom-Cole problem when $B/(-A)$ is rational. This indicates that some additional information about the continuous problem needs to be included in the numerical method. We have used the known position of the shock, x_*, to eliminate the numerically-induced shift and improve the results obtained by the inversion method. The value of x_* can be used like in [15, 25], or to divide the problem into two boundary-shock problems, but we have introduced here another possibility: our approach is based on an appropriate x_*-dependent linear combination of two numerical solutions. Whereas this resolves the difficulty with the Lagerstrom-Cole problem when $B/(-A)$ is irrational, for more complicated problems, the approach seems to be appropriate only when ε-values are fairly small. This is because x_* approximates the exact position of the interior layer better when ε is smaller.

All this shows that quasilinear interior-shock problems are indeed difficult to solve numerically.

Acknowledgments Thanks are due to two anonymous reviewers whose comments helped us improve the paper.

References

1. Allgower, E.L., Mccormickm, St.F., Pryor, D.V.: A general mesh independence principle for Newton's method applied to second order boundary value problems. Computing **23**, 233–246 (1979)
2. Bohl, E.: Finite Modelle Gewöhnlicher Randwertaufgaben. Teubner, Stuttgart (1981)
3. Chang, K.W., Howes, F.A.: Nonlinear Singular Perturbation Phenomena: Theory and Application (Applied Mathematical Sciences, vol. 56. Springer, New York (1984)
4. Farrell, P.A., Hegarty, A.F., Miller, J.J.H., O'Riordan, E., Shishkin, G.I.: Robust Computational Techniques for Boundary Layers. Chapman & Hall/CRC, Boca Raton (2000)
5. Farrell, P.A., O'Riordan, E., Shishkin, G.I.: A class of singularly perturbed quasilinear differential equations with interior layers. Math. Comput. **78**, 103–127 (2009)
6. Herman, R.L., Knickerbocker, C.J.: Numerically induced phase shift in the KdV soliton. J. Comput. Phys. **104**, 50–55 (1993)
7. Howes, F.A.: Boundary-Interior Layer interactions in nonlinear singular perturbation theory. Mem. Amer. Math. Soc. 203 (1978)
8. Kevorkian, J., Cole, J.D.: Perturbation Methods in Applied Mathematics (Applied Mathematical Sciences, vol. 34). Springer, New York (1980)
9. Lagerstrom, P.A.: Matched Asymptotic Expansions (Applied Mathematical Sciences, vol. 76). Springer, New York (1988)
10. Linß, T.: Layer-Adapted Meshes for Reaction-Convection-Diffusion Problems (Lecture Notes in Mathematics, vol. 1985). Springer, Berlin, Heidelberg (2010)
11. Lorenz, J.: Stability and monotonicity properties of stiff quasilinear boundary problems. Univ. u Novom Sadu Zb. Rad. Prirod. Mat. Fak. Ser. Mat. **12**, 151–175 (1982)
12. Lorenz, J.: Analysis of difference schemes for a stationary shock problem. SIAM J. Numer. Anal. **21**, 1038–1053 (1984)
13. Miller, J.J.H., O'Riordan, E., Shishkin, G.I.: Fitted Numerical Methods for Singularly Perturbation Problems. World Scientific, Singapore (1996)
14. O'Malley, R.E. Jr.: Singular Perturbation Methods for Ordinary Differential Equations (Applied Mathematical Sciences, vol. 89). Springer, New York (1991)
15. O'Riordan, E., Quinn, J.: Numerical Method for a Nonlinear Singularly Perturbed Interior Layer Problem. In: Clavero, C., Gracia, J.L., Lisbona, F. (eds.) Proceedings BAIL 2010 (Lecture Notes in Computational Science and Engineering 81), pp. 187–195. Springer, Berlin, Heidelberg (2011)

16. O'Riordan, E., Quinn, J.: Parameter-uniform numerical methods for some linear and nonlinear singularly perturbed convection diffusion boundary turning point problems. BIT **51**, 317–337 (2011)
17. O'Riordan, E., Quinn, J.: A singularly perturbed convection diffusion turning point problem with an interior layer. J. Comp. Meth. Appl. Math. **12**, 206–220 (2012)
18. O'Riordan, E., Quinn, J.: Numerical experiments with a Shishkin numerical method for a singularly perturbed quasilinear parabolic problem with an interior layer. In: Dimov, I., Farago, I., Vulkov, L. (eds.) Numerical Analysis and Its Applications: 5th International Conference, NAA 2012, Lozenetz, Bulgaria, June 15–20, 2012, Revised Selected Papers, (Lecture Notes in Computer Science 8236), pp. 420–427. Springer, Heidelberg (2013)
19. Osher, S.: Nonlinear singular perturbation problems and one sided difference schemes. SIAM J. Numer. Anal. **18**, 129–144 (1981)
20. Roos, H.-G., Stynes, M., Tobiska L.: Numerical Methods for Singularly Perturbed Differential Equations (Springer Series in Computational Mathematics, 2nd edn, vol. 24). Springer, Berlin (2008)
21. Shishkin, G.I.: Grid approximation of a singularly perturbed quasilinear equation in the presence of a transition layer. Russian Acad. Sci. Dokl. Math. **47**, 83–88 (1993)
22. Shishkin, G.I.: Difference approximation of the Dirichlet problem for a singularly perturbed quasilinear parabolic equation in the presence of a transition layer. Russian Acad. Sci. Dokl. Math. **48**, 346–352 (1994)
23. Vulanović, R.: Finite-difference schemes for quasilinear singular perturbation problems. J. Comp. Appl. Math. **26**, 345–365 (1989)
24. Vulanović, R.: Continuous and numerical analysis of a boundary shock problem. Bull. Austral. Math. Soc. **41**, 75–86 (1990)
25. Vulanović, R.: A uniform numerical method for a class of quasilinear turning point problems. In: Miller, J.J.H., Vichnevetsky, R. (eds.) Proceedings 13th IMACS World Congress on Computation and Applied Mathematics, p. 493. IMACS, Dublin (1991)
26. Vulanović, R.: Numerical methods for quadratic singular perturbation problems. Proc. Dynamic Systems and Appl. **2**, 543–551 (1996)
27. Vulanović, R.: Boundary shock problems and singularly perturbed Riccati equations. In: Hegarty, A.F., Kopteva, N., O'Riordan, E., Stynes, M. (eds.) Proceedings BAIL 2008 (Lecture Notes in Computational Science and Engineering 69), pp. 277–285. Springer, Berlin, Heidelberg (2009)
28. Zadorin, A.I.: Existence and uniqueness of the solution of certain difference problems for a quasilinear ordinary differential equation with a small parameter. (Russian) Chisl. Metody Mekh. Sploshn. Sredy **15**, 33–44 (1984)